South Carolina Standard Setting Study 2: Literature Review

Presented to:

South Carolina Education Oversight Committee

Presented by:

Computerized Assessments and Learning, LLC

June 27, 2009

Optional Study 2: Comparative Study of the PASS Vertical Score Scale and a System of Vertically Moderated Standards

As per the RFP requirements, Optional Study 2 is to produce a review of the literature with results that will help inform decisions on appropriate procedures and methodologies addressing the measurement of student growth within the context of South Carolina's educational accountability program. Targeted in the review are to be studies that will assist in providing information to decide on the appropriateness of developing PASS vertical score scales and/or implementing a system of vertically moderated standards. The outcome of this literature review is intended to provide recommendations delineating the relative advantages and disadvantages of the vertical score scale, vertically moderated performance standards, or a combination of both measures for calculating school and school district growth ratings. This report also provides a glossary of non-technical definitions of key terms prior to the references.

The adequate yearly progress (AYP) provision of the federal No Child Left Behind Act of 2001 (NCLB) requires every state to have annual assessments in math and reading for grade 3 through 8 and once in high school, as well as science assessment administration once during grades 3-5, 6-9, and 10-12. In addition to being required to demonstrate progress as part of the AYP process, South Carolina requires an annual assessment of achievement growth by groups of students as part of its school report card growth rating reporting. The expectation requires data based on longitudinally matched individual student data as described in the *2007–2008 Accountability Manual* provided by the South Carolina Education Oversight Committee (EOC). We begin this paper with an overview of the methods and approaches associated with Study 2.

OVERVIEW

Measuring performance over time is a complex problem. If we could give the same or parallel form of the same test over the period, it would be a straightforward solution. However, in education settings when "tests" are separated by a year or more and time is of the essence, administering the same test twice make little sense. For example, sixth grade students need to be evaluated on the sixth grade content they were expected to

learn, whereas seventh grade students will be assessed over their seventh grade material. The need to measure, track and report student progress (growth) necessitates examining procedures for studying and then linking measurements across time, grades or occasions with the intent of providing information on the developmental change in student learning and achievement. Linking methodologies referenced and discussed in the literature to achieve this purpose include procedures for equating (horizontal and vertical), vertical scaling/linking, vertically moderated standards setting (VMSS) and other growth targeted models. These procedures are fundamentally different in many ways including purpose, each relying on differing assumptions that must be considered, methodological choices needing to be made, and criteria that need to be evaluated prior to deciding which linking method to utilize. The strengths and weaknesses of each procedure need to be examined relative to purpose, to ensure the chosen method will provide the most accurate and meaningful results, with the least amount of misinterpretation or error.

It is important to clarify what linking methods may be used by test developers and state education departments to obtain a scale with properties they require. The need for comparisons between the results of one test to the results of another requires the examination of the nature of the link needed, the inferences that will be drawn from the comparison, and the degree of precision desired (Linn, 1993). A conceptual framework for linking was provided by Mislevy (1992) and Linn (1993) that included four types of statistical linking: equating, calibration, statistical moderation and projection or prediction. They also discussed social moderation which was based on judgmental procedures rather than on a statistical process.

Equating is the most powerful statistical procedure that can be used to link scores on test forms, but the adequacy of results depend on a set of severe conditions or underlying assumptions. Scores obtained from equated forms are interchangeable, and therefore, result in the strongest link for two tests; the higher the degrees of similarity, the stronger the interchangeability of the scores. However, equating is only appropriate when forms are identical in content and structure; therefore, tests that contain different levels of difficulty, or utilize variations in content and/or structure cannot be truly equated. The latter is likely to be the case for test forms developed to measure outcomes at different grade levels.

Vertical scaling or vertical linking is designed to provide a continuous scale, across

all grades, and is tolerable to changes in content and difficulty. In this respect, vertical linking could fall under procedures for calibration, statistical moderation or projection within the Mislevy and Linn framework, which type dependent on the degree and amount of content and construct overlap and the degree of psychometric similarities across grade level tests. It is important to note that though vertical scaling provides scores that are mapped onto a single scale, the scores are not equated because the method is used to assess content that changes over grades (Kolen & Tong, 2009). This change in content across grades negates the ability to use equating procedures to link the test scores.

In contrast to use of statistical procedures mentioned above as the primary methods to link scores, vertically moderated standards focus predominantly on linking student performance categories. This method can provide critical information on how the instituted standards and curriculum are working, as well as provide information about the progression of achievement through the grades; however, this method cannot provide interpretations of student growth per se (Kolen & Tong, 2009). This method is predictive in nature, theoretically allowing a state to determine a student's likelihood of meeting proficiency in a subsequent grade based on the student's performance on the earlier grade assessment.

Other growth oriented models vary in the degree to which they require statistical linking of test scores across grade levels. A paper commissioned by the Council of Chief State School Officers (CCSSO, 2008) attempts to provide a guide for those considering the implementation of a growth model to address school accountability. They identify five accountability model types: Status Models, Improvement Models, true Growth Models, Value-added Models and Transition Matrix Models. Vertical linking of test scores across grade levels is required for three of the models/methods: Improvement, true Growth and Value-added, but not for the Status or the Transition Matrix models. Please note that the Transition Matrix model is the approach currently used in South Carolina. The Transition Model requires tracking of students across grade levels, but inferences on growth are made relative to performance categories of proficiency (similar to a VMSS approach). In the selection of any approach or combination of approaches, the purpose, interpretation intent and focus (student or school) need to be the initial primary consideration. In the next sections, each of these categories and associated methods with specific applications is considered and discussed.

Test Score Equating

Psychometrically, equating is the most understood method (see Holland & Dorans, 2006 and Kolen & Brennan, 2004 for methods). Traditional test equating requires that test forms be identical; essentially, the two tests are designed to be the same in form and content, and the resulting number correct scores are indistinguishable. The tests that are to be equated must measure the same construct, and must do so at the same level of difficulty, at the same level of accuracy, and be population invariant. The tests need to have equal reliability, the equating function for equating test one to test two must be the inverse of the equating function for equating test two to test one, and test equating should not be time dependent. If these are not characteristics of the two tests, the implementation of equating procedures will not result in test scores that have the characteristics and interpretations of equated scores. If the tests are designed in such a manner, equating results in equal interval scores, placed on a single scale, and allows the scores from one form to be interpreted as identical to the scores from the other form. Horizontal and vertical equating are two methods used to equate different tests.

Horizontal Equating

Horizontal equating occurs within grade and is designed to assess students that are assumed to be at approximately the same knowledge level. The conditions required for equating are more likely to be met for the horizontal equating of within grade level test form scores. A variety of designs and methods based on both classical and Item Response Theory (IRT) procedures are available to conduct these studies (Holland & Dorans, 2006; Kolen & Brennan, 2004). When conditions are met, horizontal equating provides scores that are interchangeable, because the tests are essentially one test. One limitation of this form of equating is that it is only designed for within-grade applications and will not provide educators with information regarding student growth over time.

Horizontal equating supports the utilization of student growth percentiles and growth trajectory analyses. Growth percentile results provide states with information pertaining to the amount of growth a student has exhibited relative to their peers, while growth trajectory models study growth percentile results to formulate how much growth a student needs to have to reach or

maintain proficient status (Betebenner, 2008). In Betebenner's study, he examined data from the South Carolina Palmetto Achievement Challenge Test (PACT) to investigate the appropriateness of the information provided by growth percentiles regarding student progress. His analyses and results demonstrated that students classified as proficient were more likely to maintain that status than were non-proficient students to reach proficient. The information provided by growth percentile and growth trajectory analyses could supply states with the information needed to establish the achievement target of universal proficiency by highlighting the amount of growth each student needs to achieve or maintain in order to be classified proficient.

Vertical Equating

Vertical equating is used to assess students who are at different levels of education, utilizing the same content, and is typically used at adjacent grade levels, when content is more closely related, providing scores that are directly comparable on a continuous scale or dimension (Holland & Dorans, 2006). Unlike horizontal equating, which allows for the comparison of students within a grade, vertical equating attempts to develop a unidimensional scale, summarizing student achievement for direct comparison across grade levels (Lissitz & Huynh, 2003). The most useful method for measuring student growth across all grades would be vertically equated scores, scaled to yield equal interval scores (Clemans, 1993).

Vertical equating is distinguished from vertical scaling or linking because the content being assessed for vertical equating is assumed to be the same content across grade level test forms. The underlying assumption requirements for vertical equating are the same as for horizontal equating, i.e., tests are parallel in content and technical structure (Patz, 2007). It might be argued that vertical equating is most appropriate when assessing progress at adjacent grades; it likely will always be problematic when there is a need to assess students across grade levels or at large grade spans, such as assessing progress at grade 3 and at grade 8, because the type of assessment and content could be vastly different.

Though equating provides the strongest statistical links of separate tests, it is difficult to meet all of the criteria required for true equating conditions; consequently, many test

developers and state departments are shifting their focus to less stringent psychometric procedures in an attempt to meet the requirement of tracking student growth across time.

Vertical Scaling

Vertical scaling includes procedures that attempt to establish a continuum of student growth across grade levels when specific content dissimilarity is known to exist, but the primary dimensionality of the construct domain being measured remains intact. The scales are designed and intended to allow for direct comparison of student growth across grade, and provide districts with the ability to track and monitor student progress from year to year (Kolen, 2003; 2004; Lissitz & Huynh, 2003; Patz, 2007). Vertical scales are considered developmental because they focus on the changes in a student's scores across grades, highlighting the importance of monitoring changes in the student's competency level (Kolen, 2006).

In contrast to equating, the reduced assumptions underlying the development of vertical scaling allows states to link test scores despite having some degree of differing content, constructs, and difficulty level. When two or more tests are intended to be a representation of a developmental continuum for a subject, and the scores are functionally related, meaning the scores from Test One and Test Two are designed to result in relatively equal score probabilities, vertical scaling is an appropriate model (Harris, 1991). Tests that contain homogenous subject matter and include cognitive dimensions that are similar at every grade level foster the ability for test developers to vertically link the two forms. Vertical scaling allows researchers to link tests that are intentionally different in difficulty across grade levels (Huynh & Schneider, 2005; Lissitz & Huynh, 2003); however, the resulting links are population dependent and require the evaluation of the similarity of the tests and the appropriateness of the person or group for whom the scale is to be used (Kolen, 2004). When utilizing vertical scaling, states must determine that the two tests they wish to link are structurally similar, though some degree of divergence in content and difficulty is permissible, and must be aware that the comparisons information resulting from the vertically linked tests are not intended for use in comparing groups outside the intended linking. For example, linking third grade and fourth grade tests would provide information about third and fourth grade students and should not be used to compare outside those grades. Because vertical scaling maps

student scores on an overall scale, it produces indirect links between the scores of different levels of a test with the intention of being able to compare performance across time and student group (Kolen, 2006).

It is important to note, however, that vertical scaling provides the most meaningful information when the content standards are vertically aligned, with considerable grade-to-grade overlap and increases in difficulty, and there is a robust design defining the data collection and the psychometric procedures that will be used (Patz, 2007). Once states have established they want to implement a vertical scale, assumptions, methodology, and the method's strengths and weaknesses again need to be evaluated to determine the ability of the scale to provide the type of interpretations the state needs. Vertical scaling is much more complex than equating, and the scale can be substantially affected by the choices states make with regard to methodology and procedure; therefore, careful consideration is warranted (Kolen, 2003, 2004; Tong & Kolen, 2007).

Assumptions

To properly construct a vertical scale, underlying assumptions must be acknowledged during development and implementation. There is an equal interval assumption that needs to be addressed. The benefit of vertical scaling is supported by the scale's ability to allow for comparisons of growth at different grades for a student, or for comparison of different groups of students at different grades; however, if the equal interval assumption is not met, such comparisons are difficult (Lissitz & Huynh, 2003; Schaefer, 2006) and interpretations are likely to be flawed. To link across grades, test forms should contain comparable content and the same dimension or dimensions should be assessed in each grade, implying that teacher instructional efforts in each grade also be comparable. The degree to which the content dimensions assumption is met will impact the quality and meaningfulness of the vertically linked scores' intended use across grades.

Data Collection

To establish the vertical scale, several data collection designs have been developed. These designs include a single group design, calibration, equivalent groups, scaling test, anchor test, and common item designs (Holland & Dorans, 2006; Kolen, 2006; Kolen &

Tong, 2009). Each of these designs employs different techniques that will result in a common scale of scores; however, states and test developers would need to examine the benefits of each to determine the appropriate design to execute. Again, the scale will be affected by the method and procedures used to create it; therefore, researchers should thoroughly examine the benefits and anticipated interpretations of the scale to guide the decision making process (Tong & Kolen, 2007, 2008).

Single Group Design

Single group designs are relatively straightforward in their definition and application. This design requires a single sample from a specified population take both tests whose scores are to be linked (Holland & Dorans, 2006; Kolen & Tong, 2009). Because examinees have taken both tests, this design directly controls for variations in examinee proficiency. One major advantage of single group designs is the small sample size required.

This design is potentially subject to order effects. This can be addressed by randomly assigning the order in which the two tests are administered. While this makes this design more difficult to implement, it is critical to the quality, fidelity and accuracy of the linking results.

Calibration

Researchers use calibration when there is a need to compare scores from one test to those on another test. Developers design the tests with the same framework, but utilize different test specifications (Kolen, 2004). To ensure a calibration adequately determines the performance of students or the percentage of students scoring above a specified level, it is imperative test designers match the forms with regard to content, cognitive demands, and administration conditions (Linn, 1993). Calibration is different from equating in that it allows for performance assessment at different levels or with different degrees of reliability. Because calibration involves relaxed requirements, there is a need for different conversions of estimates for individual and for group distribution characteristics, and there is a need to check the stability over contexts, groups, and time. When linking scores across grades, the calibration process allows developers to place scores for all grades levels on the same developmental score scale (Kolen & Tong, 2009).

Equivalent Groups Design

This design uses randomly equivalent groups of examinees to take either a grade-level test or a test designed for an adjacent grade, generally from the lower grade (Kolen & Tong, 2009). Using the equivalent groups design is consistent with the grade-to-grade definition of growth. In this design, there are equivalent examinees taking all versions of the test; therefore, it is theorized their scores may be linked and used for comparison. This design does require a special administration; therefore, states should examine the administration process prior to deciding to use it. Though the theory of equivalent groups is valid, it can be difficult to implement. Two ways to address this concern include randomly selecting the student groups from a population and by testing the samples after spiraling the two tests together (Holland & Dorans, 2006). This design avoids the issues of order effects noted with single group designs because the examinees are only taking one of the tests.

Scaling Test

The scaling test design requires developers to design a separate scaling test that covers the entire content domain and is developed to be administered in a single session (Kolen, 2003; Kolen & Tong, 2009). Students from all grades take the scaling test as well as a grade appropriate test and are instructed to do their best and not guess on items. This method is closely aligned with the domain definition of growth and produces a vertical scale through utilization of the scaling test scores. Developers can develop the scale because all examinees have taken the scaling test, facilitating the linking process. The scaling test scores are used to develop the scale to measure growth over grades and the grade level tests are statistically linked to the scale. Research has shown that scaling test designs tend to produce results that are different from equivalent groups and common item designs (Kolen & Tong, 2009).

Anchor Tests

States can make use of anchoring through the development of a separate external anchor test. With an external anchor test, one group of students take the anchor test and test one, while another equivalent group takes the anchor test and test two. Test

developers then use the anchor test to adjust for any differences noted in the proficiency of the samples (Linn, 1993). Ideally, the anchor test should have a strong relation to both test forms; however, finding a stronger relation to one form provides an indication that the two tests may not be measuring the same thing (DePascale, 2006). In the situation where test developers need to link a new test to an older, no longer used test, the utilization of an anchor test allows the developers to tie the scores of the two tests (Livingston, Dorans & Wright, 1990). The use of an anchor test utilizes the scores as a stratifying variable to match the old form to the new form.

Common Item Design

Common item designs utilize the over lapping items within the two tests to be linked; the performance on these common items then serves to form the vertical scale. In this method, a test designed for grade 3 will contain items that are also on the test for grade 4. Common item designs directly determine grade-to-grade growth measures and establish commonality across grades by placing different grade tests on the same scale (Kolen, 2003). This method requires developers to design grade appropriate tests with common items embedded in tests for adjacent grades and involves students taking one grade level form. When developing the tests, there should be an established definition of the content standards for the two (adjacent) grades and items selected for inclusion in both forms should be content representative of both grades (Kolen & Tong, 2009).

This design provides states with data supporting the definition of grade-to-grade growth. To provide a solid link, there must be an adequate number of anchor items (approximately 15) to establish a representation of the domain (Patz, 2007). These items need to be determined by the shared content standards between the grades. This removes the concern of whether or not the item was originally intended or is operationally used at the grade above or below any given grade. It is important to note that research has shown that embedding items that are representative of the content area and paying particular attention to the difficulty level should improve the interpretations provided by the vertical scale (DePascale, 2006).

As the associated importance of the links increases, so should the number of anchor items. Including more items ensures the links are appropriate and support the resulting interpretations.

With regard to the ease of implementation, the common item design can be embedded into the operational testing, making it administratively simple; however, this design is also the most prone to contextual effects (Kolen & Tong, 2009). When using the common item design there is no need for developers to conduct separate research to create the vertical scale, data resulting from the operational test administration will support the construction of the scale.

Vertical Scaling Methodology

When developers are constructing the vertical scale, the obtained score on the test(s) is related to a single interim scale which is transformed to a vertical scale through statistical procedures (Kolen, 2006). Tong and Kolen (2008) noted that the common implementation of vertical scales includes constructing the vertical scale, then maintaining it through horizontal equating after the base year as new grade level forms are introduced. This can be accomplished by horizontally equating the second testing year to the base year within grade and place the two years on the base-year scale, or by developing a vertical scale for the base year and the second year data and horizontally equate the two scales, placing the two scales onto the base year scale (Tong & Kolen, 2008). A variety of psychometric procedural methods are available for use, e.g., Hieronymus, Thurstone, and IRT based methods (Kolen & Tong, 2009).

When developers have used common item collection, the percentage of common items should be greater than 20% and the items should be embedded throughout the test as opposed to at the beginning or end of the test (Kolen & Tong, 2009). If common item design and separate calibration are to be used, the developers need to calibrate each level of the test, set a base grade, and link each pair of adjacent grades, allowing all grades to be mapped onto the common base scale. With common item and concurrent calibration, the developers would need to create an incomplete data matrix that includes items that are not completed by the student coded as items not reached and use an IRT program that allows for multiple groups calibration, the resulting scores would be on the same scale due to the calibration.

IRT based methods function differently when equivalent groups design has been chosen as the data collection method. Developers need to determine if there is a need for separate or concurrent calibration. Concurrent calibration requires calibration of all

grades levels at once, coding student responses into an incomplete data matrix, this needs only one calibration and no further linking is required (Tong & Kolen, 2008). Separate calibration requires a linking chain to place all grade levels onto one scale demanding more steps; however, separate calibration avoids multidimensionality concerns that plague concurrent calibration (Kolen & Tong, 2009). When the scaling test has been used, developers need to evaluate separate and concurrent calibration again. With separate calibration, developers need to decide if they are going to calibrate the scaling test and the level test together or calibrate them separately, and for concurrent calibration should the scaling test be the only test calibrated, or should it be calibrated with the level tests.

To provide developers and states with the most useful and stable information, calibration in and of itself should be evaluated. According to Kolen and Brennan (2004), there are some very significant differences between concurrent and separate calibration. They conclude that concurrent calibration is more efficient and uses the most amount of information in the calibration process; however, they also noted that concurrent calibration may have convergence problems and past research has highlighted a concern with multidimensionality. As for separate calibration, Kolen and Brennan noted that the common items can be compared across grade levels, and though extra linking is required, there is less of a concern with multidimensionality. A recent study by Rogers, Swaminathan and Andrada (2009) concludes that separate group calibration may be the method of choice based on their results, but note that the concurrent calibration implemented was similar to separate group scaling. There also needs to be attention paid to the goal of the calibration, including determining how many grades the resulting scale will cover. Research has shown that when the scale is meant to link adjacent grades, both separate and concurrent measures resulted in similar scales; whereas, when the scale is meant to cover a large number of grades the results of concurrent and separate calibration procedures appear to vary greatly, though Kolen and Brennan suggest using separate calibration for scales meant to cover a large grade span.

Scale interpretation

Researchers use different statistical indices to evaluate the characteristics of vertical scales. States can determine the amount of grade-to-grade growth by evaluating the mean,

median, or percentile information, or evaluate the changes in scale score variability across the grade levels (Kolen & Tong, 2009). Grade-to-grade growth has been shown characteristically to decline over grades, whereas, within grade variability has been shown to increase over grades (Kolen, 2003). This increase in variability supports the theory that higher proficiency students progress at a faster rate than the lower proficiency students. Growth trends can be evaluated by examining the effect size, which suggests the amount of growth that has occurred from one grade to the next (Kolen & Tong, 2009). This very notable tendency deserves careful reflection when later in this paper Vertically Moderated Standards and other approaches to monitoring growth are presented.

Advantages and Disadvantages

Properly designed vertical scales offer a wealth of information for monitoring student growth across grades and over time. The scale can be designed to be computationally very simple, providing straightforward links between the assessment and the accountability systems (Patz, 2007). If properly designed, the scales allow states to describe student growth relative to the construct (DePascale, 2006).

Appropriately designed vertical scales are intended to approximate equal interval scales; therefore, these scales offer districts the ability to utilize common language for discussion of student achievement across grade levels and allow for meaningful, continuous tracking of student performance (Lissitz & Huynh, 2003). By rescaling the raw scores, the scores from different tests and across time can be interpreted in the same way, and the scores reported to the public are always on the same scale (Ito, Sykes & Yao, 2008; Lissitz & Huynh, 2003). The scales determine how much growth has occurred over time and at different points on the proficiency range, which allows for comparisons of one group to another and students groups at any point (Patz, 2007). Vertical scales are beneficial for linking reading or mathematics tests because of the patterned and sequential learning throughout the schooling process; however, the scale is much more difficult with tests involving science or social studies (Lissitz & Huynh, 2003).

Though vertical scaling has a number of strengths, there are weaknesses with the design that should be considered. While not as stringent as required for vertical equating, the assumption of comparable content is still critical with vertical scaling and needs

consideration. When attempting to link forms across more than two adjacent grades, it is unclear if the same dimensions are being assessed. The items and assessment processes typically change over grades; therefore, vertical scaling confounds content changes with method changes, violating the assumption of comparable assessment across grades and makes interpretation difficult (Lissitz & Huynh, 2003; Schaefer, 2006).

There is also a concern with the ability of vertical scales to assist educators in moving children into the state defined proficiency category because vertical scales may require a student to grow more in one achievement category than the other (Huynh & Schneider, 2005). For example, a student may be required to grow or learn more in mathematics to maintain a category above proficient due to the higher cognitive abilities needed to master more difficult and cognitively challenging material, whereas, a student might need to show less growth to move from below proficient to proficient. Adjacent grade levels are also not parallel; consequently, equating is not possible, and linking provides weaker comparability of test scores than equating would because the scores are not interchangeable and the validity of the interpretations between test forms is weaker (Patz, 2007). This validity also diminishes as grade level increases, and linking achievement measures at non-adjacent grade levels is much more difficult, partially due to the lack of content standards and proficiency distribution overlap (Patz, 2007).

Because vertical scales are developed with tests designed to differ in difficulty and content, the results are limited in their interpretations; therefore, the use of a vertical scale invites misinterpretation of the results (Kolen & Tong, 2009; Schafer, 2006). The interval-level interpretations between grades are also of concern because the achievement level descriptions of what a student knows are different across grades for the identical score (Schafer, 2006). There is also apprehension in utilizing vertical scales to summarize growth across grades because it may diminish important grade-level content knowledge and skills (DePascale, 2006).

If changes in scale scores are a focus of accountability methods, it is important to note that though vertical scales are designed to approximate equal intervals, achievement scales rarely achieve truly equal interval properties, this alone warrants caution and additional validation efforts (Patz, 2007; Schaefer, 2006). It is unlikely that developers can relate the scale in a consistent manner to the amount of growth needed to achieve standards based proficiency from grade to grade (Gong et al., 2006).

Unidimensionality is rarely met in real data. Performing and learning are essentially multidimensional variables (Lissitz & Huynh, 2003). Therefore, there is a legitimate concern regarding the validity of vertical scales because it is difficult to determine the impact violating the unidimensionality assumption has on vertically scaled results.. This makes the utilization of vertical scales to provide information on these activities problematic. The tests are not parallel; therefore, the resulting scores are not interchangeable.

Creating a vertical scale is technically difficult even with IRT models. There are many decisions that must be made and after development, the results will require artificial adjustments (Kolen, 2006). The scales can be technically challenging to produce and maintain, and their use is controversial in light of the scale's construct validity (Gong et al., 2006). It is also important to note that vertical scales are sample, data collection design, and scaling method dependent: therefore, states must consider those variables when evaluating the implementation of a vertical scale (Kolen, 2003, 2004; Tong & Kolen, 2007, 2008; Kolen & Tong, 2009).

Many states use the one-parameter IRT model (Rasch) to scale their scores and equate test forms within grades (Wolf, 2004; Smith & Smith, 2004; 2007). If a particular statistical approach is used for such purposes, then it would be logical to use the same procedure for vertical linking. Of importance is that the adequacy of the linking results will need to be evaluated within the context of the statistical procedure implemented and the design used. Use of other procedures and designs are likely to produce different results of varying degree. The choice of the psychometric method should be made, then that model carried forward.

Vertically Moderated Standards Setting

Adequate vertical scales are quite complicated and problematic to produce; consequently, growth scales that do not rely on meeting vertical equating or vertical scaling assumptions have been proposed as an alternative for states needing to assess student growth or progress toward the standard (Ferrara, Johnson & Chen, 2004; Lissitz & Huynh, 2003; Schafer, 2006). Ferrara, et al proposed procedures based on vertically articulated standards. Lissitz and Huynh proposed procedures for vertically moderated standards setting (VMSS) and Schaefer proposed a method for defining a growth scale

following vertically moderated standards setting.

These methods focus on grade level performance category cut scores and are designed to provide meaningful information to states regarding the progression of standards across grades or to allow states to make reasonable predictions about a student's future performance (Cizek, 2005). State departments may choose to implement VMSS models because the method fosters the establishment of consistent performance standards that are guided by policy (Lewis & Haug, 2005). This method defines progress in terms of a student meeting year-end performance expectations that predicts the student's ability to successfully meet the challenges of the next grade level (Lissitz & Huynh, 2003).

In theory, VMSS procedures allow states to identify students who have achieved the proficient standard on one year's test score and then infer them to have made sufficient yearly progress required if they attain the proficient level in the next grade (Lissitz & Huynh, 2003). This method shifts the focus from the content to growth relative to a standard or growth relative to other students, allowing states to use common language across grades for reporting, and provide information that can be used to project a student's success in the next grade (DePascale, 2006; Lissitz & Huynh, 2003). This highlights an important difference between vertical scaling and VMSS. Vertical scaling allows a state to track student growth by placing tests from adjacent grades onto one scale, allowing for direct comparison between the two grades, whereas VMSS allows a state to predict a student's future success based on the student's current performance, but does not allow for grade-to-grade comparisons. Huynh and Schneider (2005) defined VMSS as a set of common achievement level definitions and the application of a consistent trend line imposed on the percentage of students in the performance categories. VMSS allows for the performance level descriptors to be aligned and provides information about the impact of the standards on student performance across grades (Kolen & Tong, 2009).

The VMSS design requires that states implement an additional judgmental process to the standards setting as well as a statistical process that will enable them to determine if the student will be likely to attain the proficiency level of the predetermined state categories (Lissitz & Huynh, 2003). By utilizing the judgmental process, it is theorized that the resulting cut scores will progress in difficulty from grade to grade, allowing the

scores to map onto a vertical scale (Patz, 2007). The state will need to conduct standard setting studies for each grade level and then examine the set of developed standards across all grade levels (Huynh & Schneider, 2005). VMSS in essence is a post hoc approach that is applied after the development of the content standards and the assessment to produce a consistent set of performance standards across grades (Cizek, 2005; DePascale, 2006).

Assumptions

The assumptions for VMSS designs vary from the assumptions of vertical scaling. There is an expectation of consistency in curriculum and expectations across grade levels that are fundamental in the design of VMSS methods. This assumption correlates to the amount of variability in student performance across grades for a state. Because the design is dependent on the consistency in the curriculum across grade levels, it is possible the variability in student performance across grades could be due to violating this assumption. In other words, if students are not learning the same subject matter, in the same manner, and at the same rate across the state, scores could fluctuate. VMSS also relies on the judgments of professional individuals, assuming that given a large enough group, the true cut score will lie near the resulting recommended cut scores (Lissitz & Huynh). VMSS designs provide credible results presuming states have followed sound procedures when developing their content standards and that those standards are aligned across all of the grades tested. It is also assumed that the achievement level descriptors will have a relationship to the performance of students classified at each achievement level (DePascale, 2006; Lissitz & Huynh, 2003).

Establishing Cut Scores

There is an inherent need for states to establish general definitions of achievement levels across grades, which serve as guidelines for cut score determination during the judgment process of standard setting (Huynh, Barton, Meyer, Porchea & Gallant, 2005). VMSS requires states to reexamine their established cut scores, define new cut scores, as well as implement a judgment process to the standard setting which results in across-grade alignment of the states standards.

Research has suggested developing two benchmark cut scores, the current one set on

the equipercentile model used to track student growth from the base grade, the second would be an end-goal cut score based on content considerations, representing the preferred level of student performance. States could then use these two scores to determine student growth based on the current student performance levels for accountability purposes while still being able to have results focused on attaining the long-term goal (Lewis & Haug, 2005). To determine cut scores, the standard method involves utilizing the bookmark method. This method employs judges to individually determine the point among the test items ordered by difficulty where a representative student on the edge of two levels of proficiency will perform. After group discussions and viewing actual student performance data on the test, this judgment process is repeated and the final cut scores recommendations can be determined by this second round bookmarking (Buckendahl, Huynh, Siskind, Saunders, 2005).

VMSS Implementation

Lissitz and Huynh (2003) noted the elements that states must consider when evaluating VMSS methods, including the requirement that states examine their curriculum across grades, determine smoothing procedures they will use during the analysis, and conduct annual validation studies to determine the continued use of VMSS. The annual validation studies would allow the state to address any issues that might arise regarding implementation or the operationalization of the method and to evaluate any needed changes in proficiency level. States who choose to implement VMSS methods must define cut scores for each grade so that the achievement levels are consistent across grade levels and the proportion of students in each level follows a coherent growth trend across grades. The similarity of achievement levels results in a representative interpretation of the test scores and the proportion of students ensure consistency in the normative data for all of the grades.

States would need to set standards that are representative of the desired growth from the prior grade to the expected growth of the next grade (Schaefer, 2006). To set the standards across all of the grades, states might conduct standards setting for only a portion of the grades, then use interpolation and/or extrapolation to compute the cut scores for the other grades (Lissitz & Huynh, 2003). States can use an interpolation method based on knowledge of current student progress and content standards across

grades, resulting in a method that is more directly linked to the content standards (DePascale, 2006). A smoothing procedure would then provide a consistent proportion of students in each achievement level and the procedure would supplement the professional judgment involved.

Advantages and Disadvantages

By linking achievement levels, VMSS offer states an economical advantage in that the method allows states to meet the requirements of NCLB without having to invest in formal linking of different grade level tests. There is no need to link grade levels because by linking the achievement levels the state is eliminating the need to link the assessments (DePascale, 2006). By conducting multiple standard setting studies and alignment procedures, states gain information about the organization of curriculum and consistency of instruction across grades (Buckendahl et al., 2005). Because VMSS designs utilize common achievement levels across grades, the method contains a predictive aspect that states can use to assert that a child who is proficient in one grade should be expected to be proficient at the next grade (Huynh et al, 2005). The results from VMSS designs also offer common language reporting systems for all students (Lissitz & Huynh, 2003).

VMSS establishes consistency in normative data based on the assumption that the instructional efforts and opportunities are comparable across the grades (Lissitz & Huynh, 2003); however, there is no way to be certain that this assumption is being met, nor how violation of this assumption will affect the method. There is also the possibility of misinterpreting the results. It is likely that stakeholders will assume the cut scores that have been set across grades to be similar are set at equal intervals; therefore, misconstruing results as interchangeable interpretations (Huynh & Schneider, 2005). VMSS cannot be developed without the inclusion of prior knowledge of student progress and must utilize a governing body that imposes the consistency of student percentages across levels (Huynh & Schneider). VMSS procedures rely on the judgment of individuals to determine cut scores and standards; consequently, the need arises to evaluate the validity of those judgments by comparing them to the cut scores of other assessments with similar content (Buckendahl et al., 2005). Though VMSS designs establish meaningful progressions of standards across grades, there is no explicit empirical evidence if VMSS designs are capable of being used with standard-based tests

(Cizek, 2005). As such, some empirical validation in this regard would be informative.

Implementing Vertical Scales and VMSS Measures

Vertical scales describe growth relative to the construct because they do not provide information of the student's performance relative to other student nor do the scales provide information about the student's performance relative to the standard (DePascale, 2006). States requiring information aside from individual student growth may want to consider implementing measures that provide growth relative to others and the standard in conjunction with vertical scales. States can choose to apply VMSS in combination with other scales as the method works equally well in systems that function with independent horizontal scales or those that employ vertical scales. Vertically moderated standards setting can be used in conjunction with other growth models, including vertical scales. With a statistically developed vertical scale, states can apply results from the VMSS, adding standards information to the scale (Kolen & Tong, 2009). By doing so, states would be able to compare directly student scores across grades and provide predictive information about a student's future success, establishing student growth over time, as well as the student's probable future growth. States may also utilize VMSS in conjunction with a within-grade scale, resulting in an overview of the effectiveness of the curriculum, standards, and the progression of student achievement levels across grades; however, no student level growth interpretations can be reasonably made from this combination (Kolen & Tong, 2009).

Other Growth Targeted Models

Two other specific approaches attempting to address growth are the value-added statistical model methodologies (Lissitz, 2006) and a judgment based approach involving the determination and use of value tables (Hill, 2006a). The value-added approaches attempt to use student background characteristics or prior achievement and other data as statistical control variables with the intent of identifying the specific effects on student progress at the school, program or teacher level. The main purpose is to separate non-school from school based variable effects so that progress observed can be attributed to the appropriate variables (CCSSO, 2008).

Use of the value tables approach attempts to create an accountability system that

incorporates the achievement level of students from one year to the next and applies equal value to all student growth (Hill, 2005a; Hill, 2006a). The tables establish goals that are policy driven and simplify the computations and resulting information The value tables approach attempts to provide schools and states with information pertaining to status, improvement, and if the percentage of students meeting the state standard is increasing (Gong & Hill, 2004). Within these tables, states assign a numerical value to the amount of change exhibited from one year to the next, with higher values being assigned to the more valued results (Hill, Marion, Gong, DePascale, Dunn & Simpson, 2005). A proposed benefit of utilizing value tables is the system allows a school to know exactly where they stand as far as reaching the desired proficiency level, and the focus remains on student performance levels as opposed to scaled scores (Hill et al., Hill, 2006a). The system requirements for implementing value tables are rather simplistic and very basic relative to testing. States must have annual testing at consecutive grades, have the ability to track students across years to match results student by student, must have articulated standards in place that consistently define the meaning of each performance level across grades, and policy makers must define the valued outcomes (Hill et al., 2005).

Conclusion

Vertical scales and VMSS are approaches to provide information on student growth; however, vertical scales and VMSS are designed to answer very different questions about student growth and the two are mutually exclusive (DePascale, 2006). The same can be said for the other growth oriented models that are based on statistical procedures in contrast to those based on judgmental procedures. The passing of NCLB has created a need for states to find ways to measure the alignment between their curriculum and standards across grades. Because of this need, states are continuing to implement vertical scale models despite the criticisms and questions regarding their use (Schafer. 2006; Wise & Alt, 2006).

When test scores are vertically scaled, states are able to develop a common set of standards at each grade level, yet utilize information from the other test levels as well; this in turn makes it possible for states to identify the additional information or skills that the student must master to reach proficiency in one grade beyond (Patz, 2007). Huynh

and Schneider (2005) noted that if it is not possible for a state to implement vertical scaling measures, VMSS may be used to track student growth across years. The goal of VMSS is to provide states with a system to track student achievement by creating cut scores that adequately describe proficiency in terms of a student's grade level mastery and the likelihood he or she will be proficient in the next grade (Lissitz & Huynh, 2003). Lissitz and Huynh argued that if a state is considering implementing a system that focuses on individual student progression within achievement levels in current and future grades and desires a system that will provide meaningful information on classroom instruction and teacher adaptation, then vertical scales would not be appropriate. Research has centered on personal preferences with regard to scale accuracy; however, for real tests, there is no way to evaluate the "true score" or how the scale should behave; consequently, the literature regarding the distinctions between the appropriateness of different methods is based purely on preference (Yen & Burket, 1997).

A Recommendation and a Remaining Reservation

If South Carolina chooses to monitor performance of individual students over adjacent grades and over time, we offer the following recommendation in consideration of the proceeding and in recognition of the state's desire. We recommend that both a vertical scaling **and** two separate and independent approaches to achieve vertical moderated cutscores (e.g., Angoff, Ebel, Bookmark, Contrasting Groups, etc.) be explored, considered, and then implemented. The state would have data and results from at least three sources and should then work to resolve and navigate (not average) differences among procedures toward arriving at resolution. This approach allows for divergence and assimilation, while working toward decisions, understanding there is not one truth defined by one route set forth by any single solution method. In our professional judgment, this approach acknowledges the variability of methods rather than turn a deaf ear to a most nagging debate. We naturally believe we know "truth" when we select one method; with results from multiple methods to grapple with and consider, we better understand how volatile and indeterminate the truth can be. Following this recommendation, decision makers will be better informed and enlightened.

Yet even following this, or a like recommendation, we remain less than resolute in our recommendation. If the state veers from the resulting numerical and reasoned solutions of these quantitative and rational procedures and the probable decision points (that is, cutscores), then the measurement of change across grades will not be founded in the numerical analyses being recommended. Were this to be the case, we would not have confidence in any of the methodologies we have discussed and or those we recommended. For example, were cutscores to be manipulated or forced to arrive at a preferred or acceptable result, no intended objective numerical or analytical procedure can assure the quality, accuracy, fidelity, fairness or consistency of decisions or trends regardless the cost, time or dedication on behalf of the effort. In this arena, policy formed apart from psychometric and statistical guidance will often abandon the meaning and validity of test scores or the appropriateness and accuracy of the conclusion reached. Specifically, changing a decision point to arrive at a cutscore that has appeal forfeits and corrupts the determination and evaluation of growth.

Terms with Non-Technical Definitions

This lexicon is designed to serve and offer guidance to non-technical readers.

Anchor Test. This design is used when developers want to link two test forms. The developers will design an anchor test that is relatively short but which is highly correlated and comparable to both tests they wish to link. Then a student sample consisting of two equivalent groups of students take the anchor test and one form of the test to be linked. Because all students take the anchor test, any differences in proficiency across the sample groups can be accounted for by the anchor test. One example of anchor test utilization includes liking an old test form to a new test form. In this situation an anchor test is designed and a student takes the anchor test and either the old test or the new test. Scores on the anchor test are used to match the old and the new tests.

Calibration. This data collection design requires test forms be matched with regard to test content, cognitive demands, and administration conditions. Calibration places the resulting scores on a single developmental scale, which allows developers to assess student performance at different levels or grades.

Classical Test Theory. Traditional theory that researchers use during test development and eventual item selection and test preparation. This measurement approach allows the test developer to examine item and test reliability and validity in terms of item difficulty, item discrimination and overall test quality.

Common Item Design. Data collection design that uses items in tests that are to be linked. Developers embed identical items into the two tests and the scores from these embedded items form the vertical or horizontal scale for equating. An example could include linking tests that are designed to assess reading at 6^{th} grade and 7^{th} grade. The researchers will embed items in the tests that are representative of reading in both grades paying close attention to difficulty level. The score on the embedded items provide the basis for the linking between the two grades because students in both grades were exposed to the items.

Equating. This statistical process allows for the comparison of scores on different test forms, through the use of classical test theory or item response theory. Scores from tests that have been equated can be used interchangeably, allowing officials to compare scores across grades or within a grade level. Essentially, the scores from the equated forms are equal; an increase of 10 points on one is equal to an increase in 10 points on the alternate form(s).

Equivalent Groups Design. Within this data collection procedure, the goal is to link tests at adjacent grades. Researchers utilize a sample composed of two equivalent groups of students. The students then take either a grade-level appropriate test or a test designed for an adjacent grade. Because the groups are considered to be equivalent, the scores on the two tests can be linked and compared.

Extrapolation. An estimation used by developers that is representative of a value based on the extension of a known sequence of values beyond the known area, i.e.,estimating a value at a point which is outside all of the points of a known area.

Hierarchical Linear Modeling (HLM Models). Statistical method that is a form of regression or multi-level analysis. This procedure allows developers to analyze outcome variables at different hierarchical levels. For example, developers could analyze student performance within classrooms within schools.

Hieronymus Scaling Method. Scaling method that can be conducted using data from any of the discussed data collection procedures. This method utilizes the total number-correct score for tests that have items that have only two possibilities or the total number of overall points for tests that use items that have more than two possible answers.

Horizontal Equating. Equating method that is conducted within one grade using students that are, because of their identical grade level, presumed to be at the same knowledge and ability level. When thereare multiple forms of tests at a grade, horizontally equated tests are essentially and thus assumed to be one test; therefore, the resulting scores can be considered interchangeable for comparisons within that grade level. Statistical procedures allow for differences in difficulty across tests to be made. Equated (or adjusted) scores from horizontally equated tests can be used to describe the differences or similarities in student abilities within a grade. Horizontally equated tests allow for comparisons across student groups within that grade, providing meaningful information about how particular groups within the grade are performing compared to other groups within that grade.

Interpolation. An estimation used by test developers that is a representation of a value within two known values in a sequence of values. Forming an estimate of a value with reference to known values either side of it.

Item Response Theory (IRT). Statistical theory that outlines the application of mathematical models to data from tests to measure such variables as abilities and attitudes. IRT is used in test development and provides developers with information such as the probability a student will get an item correct and how test items function at different student ability levels.

Linking. See Vertical Scaling

Regression. Statistical method used to predict one variable from information provided by one or more other variables. In educational testing, it can be used to estimate the relationship of a criterion such as student performance at one grade as related to or predictable by prior achievement, gender, attendance, etc..

Scaling Test. In this collection design, developers develop a scaling test that is designed to assess the entire content domain across grades. Students from all grades take the scaling test as well as a grade appropriate test. The resulting scores produce a continuous scale by utilizing the scaling test scores, providing information about growth over grades. The grade appropriate tests are then linked to the scale. For example, developers might develop a math test that includes the mathematical content covered in school from 3rd grade to 8th grade. The sample then takes this test as well as a test appropriate for their grade level. Because all of the students took the scaling test, their performance on that test provides the scale, and their scores on the grade-level tests can then be placed on that same scale.

Single Group Design. In this data collection design procedure, one sample takes both tests that are going to be linked vertically. This process does not require an extremely large sample and because the sample takes both tests, the single administration provides developers with data that is relatively free from error resulting from sample differences in test-taker ability.

Smoothing Procedure. Process by which developers transform and adjust data to create smooth graphical representations of the test data.

Vertical Equating. Equating method that connects test forms that contain the same content and focus across grades. The scores resulting from vertical equating are used to develop a one dimensional scale that places the scores at equal intervals; therefore, an increase at one grade level or on one test is equivalent to an equal increase at another grade or alternate form. Scores from vertically equated tests are interchangeable, and can be used to assess student achievement in different grades. For example, a vertically equated math test would allow a school to compare the performance of 2nd grade students to 3rd grade students. Comparisons across non-adjacent grades might be questioned since the greater the span in grades, the more likely the difficulty level and content will considered quite different.

Vertical Scaling/Linking. Linking procedure that allows scores from tests that are intentionally different in difficulty level to be put on a continuous scale. In essence, the scores from the test forms are mapped onto one scale, and can therefore be used to compare test performance across time and student group. The scores can be used to provide information about student achievement across grades and time, allowing the state to monitor student progress from year to year. Utilizing this form of linking allows a state to track and evaluate a student's growth as they progress through school and directly compare changes in the student's competency level from year to year. For instance and in theory, vertical scaling allows a state to track a student's reading performance annually from 3rd to 8th grade and directly compare how that student's proficiency in reading has changed annually though that grade span.

Vertically Moderated Standard Setting (VMSS). This method is used to provide information to states that allows prediction of student performance in later years. VMSS utilizes a student's test score and resulting performance standard to determine that student's probability of being below, at or above that same level in subsequent years. Essentially, the method theorizes that a student's score in the 3rd grade can be used to determine or evaluate where that student will test in the 4th grade, based on an assumption of whether or not the student has attained sufficient growth to test at proficient or higher the next year.

REFERENCES

Betebenner, Damian W. (2008). *Student Growth Percentile and Growth Projection Analysis of the Palmetto Achievement Challenge Test*. National Center for the Improvement of Educational Assessment. Report prepared for the South Carolina Department of Education, 22 pp.

Buckendahl, C. W., Huynh, H., Siskind, T. & Saunders, J. (2005). A case study of vertically moderated standard setting for a state science assessment program. *Applied Measurement in Education, 18,* 83-98.

Council of Chief State School Officers (CCSSO), (2008). *Implementer's guide to growth models*. A paper commissioned by the CCSSO Accountability Systems and Reporting State Collaborative Project. Washington, DC: Council of Chief State School Officers.

Cizek, G. (2005). Adapting testing technology to serve accountability aims: The case of vertically moderated standard setting. *Applied Measurement in Education, 18,* 1-9.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment, 1,* 329-347.

DePascale, C. A. (2006). *Measuring growth with the MCAS tests: A consideration of the vertical scales and standards*. Retrieved March 17, 2009, from http://www.nciea.org/publications/MeasuringGrowthMCASTests_CD06.pdf.

Ferrara, S., Johnson, E. & Chen, W. (2004, April). *Vertically moderated standards: Logic, procedures and likely classification accuracy of judgmentally articulated performance standards.* Paper presented at the 2004 annual meeting of the National Council on Measurement in Education, San Diego, CA.

Gong, B. & Hill, R. (2002, September). *Validity of accountability models: A practical design view*. Paper presented at the 2002 annual CRESST conference, Los Angeles, CA.

Gong, B. & Hill, R. (2004, October). *Using student growth measures in school accountability.* Paper presented at the 2004 Reidy Interactive Lecture Series, Nashua, NH.

Gong, B., Perie, M. & Dunn, J. (2006). *Using student longitudinal growth measures for school accountability under No Child Left Behind: An update to inform design decisions*. Washington, DC: Council of Chief State School Officers. Retrieved March 17, 2009, from http://www.ccsso.org/content/pdfs/BGongGrowthUpdate09180BG.doc.

Harris, D. J. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28,* 221-235.

Hill, R. (2004, September). *Explicitly valuing growth.* Paper presented at the 2004 Reidy Interactive Lecture Series, Nashua, NH.

Hill, R. (2005a, June). *Measuring student growth through value tables*. Paper presented at the CCSSO LSA conference, San Antonio, TX.

Hill, R. (2005b). Establishing a value table for Alaska. Retrieved May 5, 2009, from http://www.nciea.org/publications/AKValueTable_RH05.pdf

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2005, November). *Using value tables to explicitly value student growth.* Paper presented at the Conference on Longitudinal Modeling of Student Achievement, College Park, MD.

Hill, R. (2006a, April). *Using value table for a school-level accountability system*. Paper presented at the 2006 annual meeting of the National Council on Measurement, San Francisco, CA.

Hill, R. (2006b). *Developing a value table for Alaska's public school performance incentive program.* Retrieved May 5, 2009, from http://www.nciea.org/publications/DevValueAlaska_RH06.pdf

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: American Council on Education and Praeger Publishers.

Huynh, H., Barton, K. E., Meyer, J. P., Porchea, S. & Gallant, D. (2005). Consistency and predictive nature of vertically moderated standards for South Carolina's Palmetto Achievement Challenge tests of language arts and mathematics. *Applied Measurement in Education, 18,* 115-128.

Huynh, H. & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education, 18,* 99-113.

Ito, K., Sykes, R. C. & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21,* 187-206.

Kolen, M. J. (2003, April). *Equating and vertical scaling: Research questions*. Paper presented at the 2003 annual meeting of the National Council on Measurement in Education, Chicago.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement, 28,* 219-226.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: American Council on Education and Praeger Publishers.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating: Methods and Practices* (2nd ed.).

New York: Springer-Verlag.

Kolen, M. J. & Tong, Y. (April, 2009). *Vertical Scaling.* Training session conducted at the 2009 annual meeting of the National Council on Measurement in Education, San Diego.

Lewis, D. M. & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18,* 11-34.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83-102.

Lissitz, R. (Ed.) (2006). Longitudinal and value added models of student performance. Maple Grove, MN: JAM Press.

Lissitz, R. W. & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research, and Education, 8,* 1-10. Retrieved February 19, 2009, from http://pareonline.net/getvn.asp?v=8&n=10.

Livingston, S. A., Dorans, N. J. & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3, 73-95.

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, methods and prospects. Princeton, NJ: ETS Policy Information Center.

Patz, R. J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: Council of Chief State School Officers. Retrieved March 17, 2009, from http://www.ccsso.org/content/pdfs/VerticalScaling.pdf.

Rogers, H. J., Swaminathan, H. & Andrada, G. (2009). A comparison of IRT procedures for vertical scaling of large scale assessments. Paper presented at the 2009 annual meeting of the American Educational Research Association, San Diego, CA.

Schaefer, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment, Research and Evaluation, 11,* 1-6. Retrieved February 19, 2009 from, http://pareonline.net/pdf/v11n4.pdf.

Smith, E.V., Jr., & Smith, R.M. (Eds.) (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Smith, E.V., Jr., & Smith, R.M. (Eds.) (2007). *Rasch measurement: Advanced and specialized applications*. Maple Grove, MN: JAM Press.

Tong, Y. & Kolen, M.J. (2007). Comparisons of Methodologies and Results in Vertical Scaling for Educational Achievement Tests. *Applied Measurement in Education, 20 (2), 227-253.*

Tong, Y. & Kolen, M. J. (2008). *Maintenance of vertical scales*. Paper presented at the 2008 annual meeting of the National Council on Measurement in Education, New York City.

Wise, L. & Alt, M. (2006). *Assessing vertical alignment*. Washington DC: Council of Chief State School Officers. Retrieved March 27, 2009, from http://www.ccsso.org/publications/details.cfm?PublicationID=293.

Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In E. Smith & R. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 366-390). Maple Grove, MN: JAM Press.

Yen, W. M. & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*, 293-313.